

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-132811

(43)Date of publication of application : 10.05.2002

(51)Int.Cl.

G06F 17/30

G06F 17/27

G06F 17/28

(21)Application number : 2000-319998

(71)Applicant : NIPPON TELEGR & TELEPH CORP
<NTT>

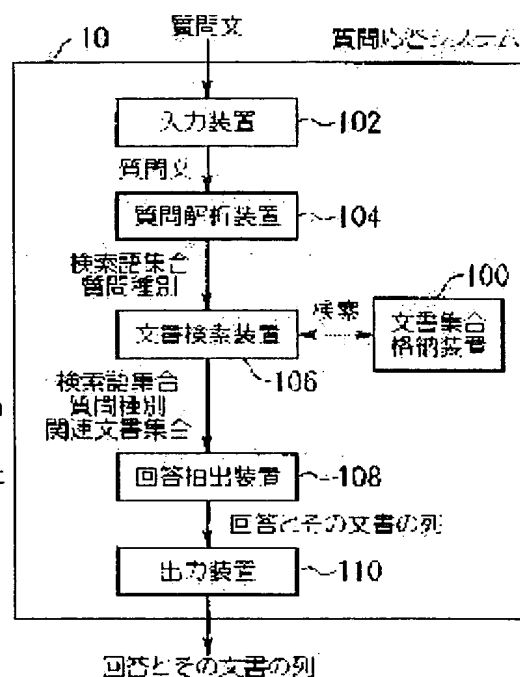
(22)Date of filing : 19.10.2000

(72)Inventor : SASAKI YUTAKA
ISOZAKI HIDEKI
TAIRA HIROYORI
KAZAWA HIDETO
HIROTA KEIICHI
NAKAJIMA HIROYUKI
HIRAO TSUTOMU
KATO TSUNEAKI(54) METHOD AND SYSTEM FOR ANSWERING QUESTION AND RECORDING MEDIUM WITH
RECORDED QUESTION ANSWERING PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To output an answer to a question sentence and its document when a document collection is given.

SOLUTION: This question answering system outputs the answer to the question sentence and its document once given the document collection and question sentence. The system has a document collection storage device 100 which stores document collections, an input device 102 which receives the question sentence, a question analyzing device 104 which decides a retrieval word collection and the kind of the question from the question sentence obtained from an input device, a document retrieving device 106 which retrieves a related document collection from the document collections stored in the document collection storage devices, an answer extracting device 108 which extracts answers to the question sentence from respective documents in the related document collection and generates an array of the answers and the documents having the answers extracted as an answer result to the question sentence, and an output device 110 which outputs the answer result.



LEGAL STATUS

[Date of request for examination]

21.12.2001

[Date of sending the examiner's decision of
rejection]

[Kind of final disposal of application other than

the examiner's decision of rejection or
application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

【請求項8】 前記回答抽出装置は、前記関連文書集合内の各文書から回答を抽出する際に、固有名称や数値表現の認識を行なうことを特徴とする請求項5乃至7のいずれかに記載の質問応答システム。

【請求項9】 文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、

質問文を受け取る第1のステップと、
入力された質問文から検索語集合と質問種別を判定する
第2のステップと、

と、
前記検索語集合および質問種別に従って、前記与えられ
た文書集合から関連文書集合を検索するステップ

と、該回答を抽出した文書の列を作成する第4のステップと、

する応答結果として出力する第5のステップと、
をコンピュータに実行させる質問応答プログラムを記録
した記録媒体。

【発明の詳細な説明】

【0001】
【発明の属する技術分野】本発明は、自然言語処理システム、質問処理システム、知識処理システム、情報検索システム、質問処理システム等、に用いられ、質問文に対する回答、情報抽出システム等に用いられ、質問文に対する回答を出力する質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体に関する。

[0002]

【従来の技術】従来の情報検索技術は、与えられた文書集合から、ユーザの入力した検索語に一致する文書の集合を取り出しても、また、従来の情報抽出技術は、与えられた文書の1つの文書の中で、分野ごとに予め決められた項目を抽出するものであった。従来の日本語質問応答システムは、回答として単語や単語の列を出力していたが、その回答の元となる記事を同時に出力していなかった。

【0003】さらに、固有名称の抽出技術の利用や、質問文に対する回答を含む可能性のある文書を検索する際の文書スコア、及び回答を抽出する際の抽出スコアの回答の文書スコア、及び回答を抽出する際の抽出スコアの回答の文書スコアを総合的に用いて、質問に対して出力する回答と文書の列の順序を決める等、回答の精度を向上させるための工夫が行なわれていなかった。

[0004]

【発明が解決しようとする課題】まず、従来の情報検索系技術は質問に対する結果を文書の単位で応答するため、ユーザが文書の内容を眺まなければならない。例えば、「日本の首相は誰ですが？」という質問に対して、返ってきた結果に含まれる文書を読むことにより初めて、文書を眺まなければならない。

【特別の文庫】

【請求項1】 文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力するコンピュータシステムを使用した質問応答方法であって、

入力された質問文から検索語集合と質問語列を判定し、該検索語集合および該質問語列に従って、前記与えられた文書集合から関連文書集合を探索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した各文書の列を前記質問文に對する応答結果として出力するところを特徴とする質問応答方法。

【請求項2】 前記関連文書集合の要素は文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする請求項1に記載の質問応答方法。

【請求項3】 前記関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに基づいて、回答と文書の列を順序付けることとを特徴とする請求項1または2のいずれかに記載の質問応答方法。

【請求項4】 前記関連文書集の各文書から回答を抽出する際に、固有名称や数値表現の認識を行なうことを特徴とする請求項1乃至3のいずれかに記載の質問応答方法。

【請求項5】 文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答システムであって、

文書集合を格納する文書集合格納装置と、
質問文を受けとる入力装置と、
格納入力装置から得られた前記質問文から検索語集合と質問語別を判定する質問解析装置と、

前記配係系集集合と質問群別に依って、前記文書集合格を格納する。
前記文書集合と、
前記文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答から抽出した文書の列を前記質問文に
対する応答結果として作成する回答抽出装置と、
前記応答結果を出力する出力装置と。

【請求項6】 前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする請求

【請求項7】 前記回答抽出装置は、前記文書検索装置が前記文書検索装置が関連文書集合を探索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書群の中から各文書から回答を抽出する際に計算した抽出スコアとを照合し、抽出スコアが前記文書スコアに比べて高い文書を前記回答とすることを特徴とする請求項5または6のいずれか1に記載の装置。

(19) 日本國特許庁 (J P)
(12) 公開特許公報 (A)
(11) 特許公開公報番号
特開2002-132811
(P2002-132811A)
平成14年5月10日(2002.5.10)
/ (3) 公開日

(51)Int.Cl.	國際記号	P I	5775-1' (參考)
G 06 F 17/30	330	G 06 F 17/30	330C 5B075
	170		170A 5B091
17/27		17/27	E
17/208		17/208	T

審査請求 有 請求項の数9 OL (全10頁)

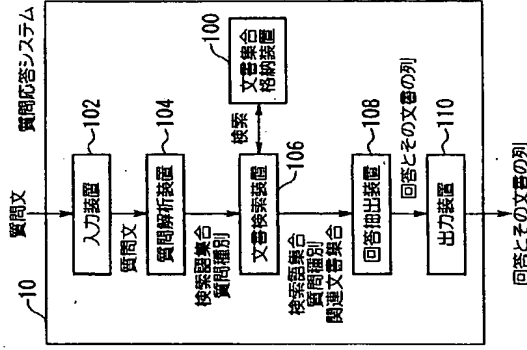
(21) 出願番号	特願2000-319998(P2000-319998)	(71) 出願人	0000044226	日本電信電話株式会社	日
(22) 出願日	平成12年10月19日(2000.10.19)	(72) 発明者	佐々木 裕	東京都千代田区大手町二丁目3番1号	日
		(72) 発明者	藤崎 秀樹	東京都千代田区大手町二丁目3番1号	日
		(74) 代理人	100054908	本電信電話株式会社	日

(54) 発明の名称] 質問応答システム及び質問応答プログラムを記録した記録媒体

(57) 【應約】

【課題】 文書集合が与えられると質問文に対する回答とその文書を出力できるようにする。

【解説手段】 文書集合と質問文が与えられると、質問文に対する回答と文書の列を出力する質問応答システムであって、文書集合を格納する文書集合格納装置 100 であって、質問文を受け入れる入力装置 102 と、入力装置から得られる入力質問文と検索結果集合と質問種別に基づいて、文書集合格納装置 104 と、前記検索結果集合と質問種別に従って、文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置 106 と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を質問文に対する応答結果として作成する回答抽出装置 108 と、応答結果を出力する出力装置 110 とを有する。



いことは「新着相」といった直接的な回答が欲しいユーザにとっては非常に煩わしいという問題が有った。

【0005】次に、従来の情報抽出技術は、予め決められた特定の分野でしか使えないため、任意の質問文に対する回答を出力する質問応答システムでは使えない。従来の日本語質問応答システムは、回答として単語や単語の列を出力していたが、回答の元となる記事を選んでいるにもかかわらず、回答を適切なユーザが得ることができなかった。例えば、「シンパゾエの大統領は誰ですか?」という質問に対して、「ムガベ大統領」と回答が返ってきた。ユーザが本当に「ムガベ大統領」が正しいかどうかを確認できないという問題が生じた。

【0006】本発明はこのような事情に鑑みてなされたものであり、質問文に対する直接的な回答とその回答の元となる文書を表示することにより回答の信頼性をユーザが確認することができ、質問応答システム、質問応答プログラム及び質問応答プログラムを記録した記録媒体を提供することを目的とする。

【0007】

【課題を解決するための手段】上記目的を達成するため、請求項1に記載の発明は、文書集合と質問文が与えられ、該質問文に対する回答と文書の列を出力するコンピュータシステムを使用した質問応答方法であって、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力することと特徴とする。

【0008】請求項1に記載の発明によれば、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力することにより、質問文に対する直接的な回答とその回答の元となる文書を表示することにより回答の信頼性をユーザが確認することができる。

【0009】また請求項2に記載の発明は、請求項1に記載の質問応答方法において、前記関連文書集合の要素は、文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することと特徴とする。

【0010】請求項2に記載の発明によれば、請求項1に記載の質問応答方法において、前記関連文書集合の要素は文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索するようにしたので、直接的な回答が得られる。

【0011】また、請求項3に記載の発明は、請求項1またはそのいずれかに記載の質問応答方法において、前記関連文書集合を検索する際に計算した各文書のスコア

である文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに従って、回答と文書の列を順序付けることを特徴とする。

【0012】請求項3に記載の発明によれば、請求項1または2のいずれかに記載の質問応答方法において、前記関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の信頼性の向上が図れる。

【0013】また、請求項4に記載の発明は、請求項1乃至3のいずれかに記載の質問応答方法において、前記関連文書集合の各文書から回答を抽出する際に、固有名称や数値表現の認識を行なうことを特徴とする。

【0014】請求項4に記載の発明によれば、請求項1乃至3のいずれかに記載の質問応答方法において、前記関連文書集合の各文書から回答を抽出する際に、固有名称や数値表現の認識を行なうようにしたので、質問文に対する精度の向上が図れる。

【0015】また、請求項5に記載の発明は、文書集合と質問文が与えられ、該質問文に対する回答と文書の列を出力する質問応答システムであって、文書集合を格納する文書集合格納装置と、質問文を受けとる入力装置と、該入力装置から得られた前記質問文から検索語集合と質問種別を判定する質問解析装置と、前記検索語集合と質問種別に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として作成する回答抽出装置と、前記文書集合中の各文書から質問文に対する回答と文書の列を前記質問文に対する応答結果として作成する出力装置とを有することを特徴とする。

【0016】請求項5に記載の発明によれば、文書集合を格納する文書集合格納装置と、質問文を受けとる入力装置と、該入力装置から得られた前記質問文から検索語集合と質問種別を判定する質問解析装置と、前記検索語集合と質問種別に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として作成する回答抽出装置と、前記質問文に対する出力装置とを有するので、質問文に対する直接的な回答とその回答の元となる文書を表示することにより回答の信頼性をユーザが確認することができる。

【0017】また、請求項6に記載の発明は、請求項5に記載の質問応答システムにおいて、前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする。

【0018】請求項6に記載の発明によれば、請求項5に記載の質問応答システムにおいて、前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索するので、直接的な回答が得られる。

【0019】また、請求項7に記載の発明は、請求項5または6のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに従って、回答と文書の列を順序付けることを特徴とする。

【0020】請求項7に記載の発明によれば、請求項5または6のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【0021】また、請求項8に記載の発明は、請求項5乃至7のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名称や数値表現の認識を行なうことを特徴とする。

【0022】請求項8に記載の発明によれば、請求項5乃至7のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名称や数値表現の認識を行なうので、質問文に対する回答の精度の向上が図れる。

【0023】また、請求項9に記載の発明は、文書集合と質問文が与えられ、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第1のステップと、入力された質問文から検索語集合と質問種別を判定する第2のステップと、前記検索語集合および該質問種別に従って、前記文書集合から関連文書集合を検索する第3のステップと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第4のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する出力装置として出力する第5のステップとをコンピュータに実行させる質問応答プログラムを記録した記録媒体を要旨とする。

【0024】請求項9に記載の発明によれば、文書集合と質問文が与えられ、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第1のステップと、入力された

質問文から検索語集合と質問種別を判定する第2のステップと、前記検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索する第3のステップと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第4のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第5のステップとをコンピュータに実行させる質問応答プログラムを記録した記録媒体に記録したので、この記録媒体に記録した質問応答プログラムをコンピュータシステムに読み込ませ、実行することにより、質問文に対する直接的な回答とその回答の元となる文書を表示することにより回答の信頼性をユーザが確認することができる。

【0025】

【発明の実施の形態】以下、本発明の実施の形態を、図面を参照して詳細に説明する。図1に本発明の装置の形態に係る質問応答システムの構成を示す。本発明の装置の形態に係る質問応答システムは、質問文に対する直接的な回答として文書の一部を取り出すとともに、回答を取り出した文書をユーザに出力することにより、回答の信頼性をユーザが確認できるようにしている。

【0026】また、固有名称の抽出技術や、与えられた文書集合から関連文書を検索する際に必要な文書スコアおよび関連文書集合から回答を抽出する際における抽出スコアの両方を総合的に用いて評価することにより、出力する回答と文書の列の順序を決めることにより、質問文に対する正しい回答と回答と文書の列のより上位に並べることができる。

【0027】すなわち、本発明の装置の形態に係る質問応答システムは、文書集合と質問文が与えられ、該質問文に対する回答と文書の列を出力するコンピュータシステムを使用した質問応答方法であって、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力することを特徴とする質問応答方法を実施するための装置である。

【0028】図1において、本発明に係る質問応答システム10は、与えられた文書集合を格納する文書集合格納装置100と、質問文を受け取る入力装置102と、質問文を解析し、上記質問文から検索語集合と質問種別を判定する質問解析装置104と、検索語集合と質問種別に従って文書集合格納装置100に格納されている文書集合から関連文書集合を検索する文書検索装置106と、関連文書集合から上記質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を上記質問文に対する応答結果として作成する回答抽出装置108と、上記応答結果を出力する出力装置110とを有している。

【0029】上記構成からなる本実施の形態に係る質問応答システム100の処理内容を図2に示すフローチャートに基いて説明する。まず、文書集合格納装置100に、与えられた文書の集合が格納される(ステップ200)。入力装置102は、質問文が入力されると、その質問文を質問解析装置104に送す(ステップ201)。質問解析装置104は質問文に対して形態素解析を行い、質問文から後述語集合と質問種別を判定し、後述語集合と質問種別を文書検索装置106に送す(ステップ202)。

【0030】文書検索装置106は質問解析装置104から受け取った後述語集合と質問種別に従って、文書集合格納装置2に格納された文書集合を後述語、関連文書集合を抽出し、辞書型文書集合の各文書に対して後述語が含まれている度合いを示す文書スコアを算出し、上記関連文書集合を、後述語集合及び質問種別と共に回答抽出装置108へ送す(ステップ203)。

【0031】回答抽出装置108は文書後述語装置106から受け取った関連文書集合中の各文書について形態素解析を行う(ステップ204)。そして回答抽出装置108は、形態素解析を行った各文書に対して質問種別に従った単語を抽出対象とし、抽出すると共に(ステップ205)、上記抽出対象についてその抽出対象が含まれる文書中における後述語との距離に基づいて抽出スコア

文書番号	文書
D1	【日本の森音相とアトランティク大断層が会談した】
D2	【ロシアのアナチン大断層が来日した】
D3	【昨日、京動で笹岡が引かれた】

本発明の実施の形態では、説明を容易にするため、文書を1文だけで記述しているが、複数の文書からなる文書でもよい。また、質問文を1つ受けて、回答とその文書の列を返す例を述べているが、これを繰り返すことにより、質問と回答を繰り返すことができる。

【0035】まず、文書集合格納装置100に文書D1、D2、D3を格納する。ここでは、表形式で表現しているが、格納の方法は文書が格納できればリストやデータベースなどの方法であつても何でもよい。以下、表形式でデータを表した例には同様のことが言える。

【0036】入力装置102は、質問文を受けとり、質問解析装置104へ送す。まず、質問解析装置104は、質問文の形態素解析を判定する。本実施の形態では、質問文の人名を聞くwho、場所を聞くwhere、日時を聞くwhenの3種類とする。なお、この他の物の名前を聞くwhatや、方法を聞くhowなどの質問種別があつたとしても、同様な方法で質問種別を判定することができる。

岩波書店、1996)に述べられている。形態素解析は、辞書に含まれる単語の列に現れるかどうかを調べ、文を辞書にある単語の列に分割し、辞書に書かれている各単語の品詞のうち、前後の単語の關係から最適な品詞を選択することにより実現される。

【0039】現在の例では、「アメリカの大統領は誰ですか?」は、表3のように形態素解析されるとする。

【表3】

Q1の形態素解析結果

単語番号	単語	品詞
1	アメリカ	固有名称
2	の	格助詞
3	大統領	普通名称
4	は	助助詞
5	誰	代名称
6	です	助動詞
7	か	終助詞
8	?	記号

KW=[アメリカ、大統領]

と替ける。ここで、自立語のみを後述語集合とするのは、説明の簡便化と後述語の向上を図るためであるので、すべての単語を後述語集合にしたり、他の選択法によって後述語集合を選択したりしても構わない。

【0041】質問解析装置104は後述語集合KWと質問種別QTを文書後述語装置106に送す。文書後述語装置106は、後述語集合KWがより多く含まれる文書を文書集合格納装置100に格納されている文書集合から探す。後述語集合の含まれている数に数え、それを各文書(文書番号で表される。)の文書スコアとする。この結果を表4に示す。

【表4】

文書スコア

文書番号	文書スコア
D1	2
D2	1
D3	0

$$RD = [D1, D2]$$

となる。なお、文書スコアの計算法はIDF法やTF-IDF法など、当業界において用いられる方法ならなんでもよい。IDF法やTF-IDF法の計算式や後述語の高速化のためのインデックス作成法は、例えば、文獻(徳永健伸:情報検索と言語処理、東京大学出版会、1999年)に述べられている。

【0044】また、ここでは文書スコアの計算には、各文書中に後述語集合の要素が現れるかどうかを文書中の

なお、ここで品詞名は言語学や自然言語処理において利用されるものであれば何でもよい。例えば、普通名詞を一般名詞と表現してもよい。このうち、普通名詞、固有名称などの自立語を後述語集合KWとする。

【0040】上記質問文から後述語集合KWは、

(1)

【0042】文書番号D1の文書には「アメリカ」、「大統領」の2つの後述語集合の要素が含まれるので、文書番号D1の文書の文書スコアは2となる。また、文書番号D2の文書スコアは、後述語集合の要素である「大統領」しか含まないので文書スコアは1となる。さらに、文書番号D3の文書は、後述語集合の要素を含まないので文書スコアは0である。

【0043】文書後述語装置106は文書スコアが0より大きい文書を関連文書集合RDとする。したがって、

(2)

文字で調べたが、文書を形態素解析し、単語に分けてから文書中の単語と後述語を比較してもよい。さらに、文書全体を1まとまりとしなくても、文書の各段落を後述語の単位として、100文字といったパッセージを後述語の単位としてもよい。文書後述語装置106は関連文書集合RDを、後述語集合及び質問種別と共に、回答抽出装置108に送す。なお、関連文書集合RDは文書全体を返してもよいし、文書の名前だけを返してもよい。

【0045】回答抽出装置108は、関連文書集合RD中の文書を形態素解析する。文書番号D1と文書番号D2の各文書の形態素解析結果をそれぞれ表5、表6に示す。

【0046】次に固有表現抽出法により、文書中の各単語が<人名>、<地名>、<日時>という3種類の固有表現のどれかに該当するかどうかを判定し、該当する場合はその種別を単語に付与する。基本的には固有表現の辞書と周囲の単語により人名、地名、日時であるかを判定する。固有表現の抽出法については、例えば特開平11-0675622に記載されている。固有表現の判定後の結果を表7、表8に示す。

【0047】

【表5】

文書D1の形態素解析結果

単語番号	単語	品詞
1	日本	固有表現
2	の	格助詞
3	森	固有表現
4	首相	普通名詞
5	と	格助詞
6	クリントン	固有表現
7	.	記号
8	アメリカ	固有表現
9	大統領	普通名詞
10	が	格助詞
11	会談	動詞
12	した	助動詞

【表6】

文書D2の形態素解析結果

単語番号	単語	品詞
1	ロシア	固有表現
2	の	格助詞
3	ブーチン	固有表現
4	大統領	普通名詞
5	が	格助詞
6	来日	動詞
7	した	助動詞

【0048】回答抽出装置108は、質問種別によって単語を抽出対象とする。whoの場合は品詞・固有表現が<人名>である単語を抽出対象とし、whereの場合は品詞・固有表現が<場所>である単語を抽出対象とし、whenの場合は品詞・固有表現が<日時>である単語を抽出対象とする。

単語を抽出対象とする。但し、質問種別に対応する抽出対象はこれに限るわけではない。例えば、whereの抽出対象に相関名を加えても良い。また、“我輩は猫である”のようなかな字括弧で括れた部分も抽出対象に加えても良い。

【0049】

【表7】

文書D1の固有表現抽出結果

単語番号	単語	品詞・固有表現
1	日本	<地名>
2	の	格助詞
3	森	<人名>
4	首相	普通名詞
5	と	格助詞
6	クリントン	<人名>
7	.	記号
8	アメリカ	<地名>
9	大統領	普通名詞
10	が	格助詞
11	会談	動詞
12	した	助動詞

【表8】

文書D2の固有表現抽出結果

単語番号	単語	品詞・固有表現
1	ロシア	<地名>
2	の	格助詞
3	ブーチン	<人名>
4	大統領	普通名詞
5	が	格助詞
6	来日	動詞
7	した	助動詞

【0050】本実施の形態の場合は、質問種別がwhoであるので、文書番号D1の文書における単語「森」、「クリントン」、文書番号D2の文書における単語「ブーチン」が抽出対象となる。これらの抽出対象について、後述語KWとの距離を使って抽出スコアを計算する。ここでは、抽出対象と後述語KWの各要素が何単語離れて出現するかを求めそれを距離とし、各要素について求められた距離の逆数の和を抽出スコアとする。

【0051】具体的には、単語間の距離は2つの単語の単語番号の差の絶対値とする。文書番号D1の文書における単語「森」は単語「アメリカ」と5単語、単語「大統領」と6単語離れているので単語「森」の抽出スコア

は1/5+1/6=0.37となる。また、文書番号D1の文書「クリントン」は「アメリカ」と2単語、「大統領」と3単語離れているので「クリントン」の抽出スコアは1/2+1/3=0.83となる。さらに、文書番号D1の文書における単語「ブーチン」は単語「大統領」と1単語離れているので単語「ブーチン」の抽出スコアは

$$\text{総合スコア} = a \times \text{文書スコア} + b \times \text{抽出スコア}$$

但し、a、bはそれぞれ文書スコア、抽出スコアの重みを表すパラメータであり任意に定め方がある。ここで、a=1、b=1とする。なお、スコアの計算法は当

$$\text{総合スコア} = a \times \text{文書スコア} + b \times \text{抽出スコア} + c \times \text{文書スコア} + d$$

×抽出スコア

のように多様化してもよい。ここで、文書スコアの「文書」は、観点で区切られた複数の文書の集合体であり、文書スコアの「文」は、観点までをひとまとまりとした文字列の集合体である。各文書中の抽出対象についての抽出スコア、文書スコア、総合スコアを計算した結果を表9に示す。

【0054】

【表9】

抽出対象のスコア

抽出対象	文書	抽出スコア	文書スコア	総合スコア
森	D1	0.37	2	2.37
クリントン	D1	0.83	2	2.83
ブーチン	D2	1	1	2

【0055】さらに、回答抽出装置108は、抽出対象の総合スコアの大きい順に抽出対象を回答とし、その文書番号を出力装置110に渡し、結果を出力する。ここでは、次のように出力される。

1. 回答 = 「クリントン」、文書 = D1

2. 回答 = 「森」、文書 = D1

3. 回答 = 「ブーチン」、文書 = D2

なお、出力の形式は同じ内容を載していれば、他の形式でもよい。このようにして、文書集合と「アメリカ大統領は誰ですか?」という入力文が与えられると、それに

対する答えを文書集合から取り出し、回答とその回答を含む文書の列として出力される。

【0056】尚、図2に示す処理内容を質問回答プログラムとして作成し、このプログラムをコンピュータにより読み取り可能な記録媒体に記録し、この記録媒体を使用することによりコンピュータシステムに質問回答プログラムを実行させることにより質問回答システムの機能を実現するようにしてもよい。

【0057】すなわち、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問回答を行うための質問回答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第1のステッ

1となる。なお、抽出スコアの計算は当業界で使われるものであれば、他のものでもよい。

【0052】次に、回答抽出装置108は、各文書中の抽出対象について、総合スコアを次のような総合計算式で求める。

$$(3)$$

業界で用いられるものであれば他の方法でも構わない。【0053】例えば、a、b、c、dをパラメータとして、

$$(4)$$

と質問種別を決定する第2のステッと、前記後述語集合および質問種別によって、前記与えられた文書集合から関連文書集合を探索する第3のステッと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第4のステッと、前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第5のステッとをコンピュータに実行させる質問回答プログラムを記録媒体に記録し、この記録媒体に記録した質問回答プログラムをコンピュータシステムに読み込ませ、実行することにより、質問回答システムの実現を実現するようにしてもよい。

【0058】なお、ここでいう「コンピュータシステム」とは、OSや周辺機器等のハードウェアを含むものとする。また、「コンピュータ読み取り可能な記録媒体」とは、フロッピー（登録商標）ディスク、光磁気ディスク、ROM、CD-ROM等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。

【0059】さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線の通信回線を経由してプログラムを送信する場合の通信媒体のように、短時間の間、動的にプログラムを保持するもの（伝送媒体ないしは伝送波）、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時間プログラムを保持しているものも含むものとする。

【0060】また上記プログラムは、前述した機能の一部を実現するためのものであってもよく、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、所謂差分ファイル（差分プログラム）であってもよい。

【0061】以上説明したように、本発明の実施の形態に係る質問回答方法、質問回答システム及び質問回答プログラムを記録した記録媒体によれば、自然言語で表現された質問文に対する回答を求めることが可能となる。また、本発明の実施の形態に係る質問回答方法、質問回答システム及び質問回答プログラムを記録した記録媒体によれば、質問文に対する直接的な回答とその回答の

とになる文書が求められるので、ユーザが直接回答を得るとともに、その回答の根拠を文書でチェックできる。

【0062】従来の情報検索技術では、文書全体の列が提示されるため、上記実施例の形態において質問文の内容である「大垣朝の名称」は、ユーザが文書を読んでも初めて見つけることができない。また、従来の情報抽出技術は分野に依存していたため、自由な質問文に対する答えを抽出することはできなかった。従来の質問応答システムは、回答とも文書を提示することができなかった。ユーザが回答の正しさを計ることができなかった。

【0063】さらに、従来の質問応答システムでは、日本文の処理に必要な単語分けを含む形態素解析を行っていないか、抽出対象の抽出を行っていないか、抽出対象のスコアを評価する際に検索時のスコアを利用したりしていないか、回答の精度が低くなっていた。

【0064】例えば、実施例において、検索スコアを無視して、総合スコアを抽出スコアとすると、回答の列の順序が、「フーテン」「クリントン」「森」の順になっている。また、形態素解析や固有表現の抽出を行わないと、抽出対象の単語が決まらず、回答の精度が低くなる。本発明の実施例の形態に係る質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体は、このような問題を解決し、質問文に対する回答とその文書の列をユーザに提示できる効果がある。

【0065】

【発明の効果】以上に説明したように、請求項1に記載の発明によれば、入力された質問文から検索結果集合と質問結果を判定し、該検索結果集合および該質問結果に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する回答結果として出力するようにしたので、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【0066】請求項2に記載の発明によれば、請求項1に記載の質問応答方法において、前記関連文書集合の要素は文書全体ではなく、文書の一部分として前記関連文書集合を検索するようにしたので、直接的な回答が得られる。

【0067】請求項3に記載の発明によれば、請求項1または2のいずれかに記載の質問応答方法において、前記関連文書集合を検索する際に計算した各文書のスコアであるスコアを前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【0068】請求項4に記載の発明によれば、請求項1乃至3のいずれかに記載の質問応答方法において、前記関連文書集合の各文書から回答を抽出する際に、固有な

や数値表現の認識を行なうようにしたので、質問文に対する精度の向上が図れる。

【0069】請求項5に記載の発明によれば、文書集合を格納する文書集合格納装置と、質問文を受けとる入力装置と、該入力装置から得られた前記質問文から検索結果集合と質問結果を判定する質問解析装置と、前記検索結果集合と質問結果に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する回答結果として作成する回答抽出装置と、前記回答結果を出力する出力装置とを有するので、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【0070】請求項6に記載の発明によれば、請求項5に記載の質問応答システムにおいて、前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部分として前記文書集合から前記関連文書集合を検索するので、直接的な回答が得られる。

【0071】請求項7に記載の発明によれば、請求項5または6のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアであるスコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの2つのスコアに基づいて、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【0072】請求項8に記載の発明によれば、請求項5乃至7のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有な単語や数値表現の認識を行なうので、質問文に対する回答の精度の向上が図れる。請求項8に記載の発明によれば、請求項5乃至7のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有な単語や数値表現の認識を行なうので、質問文に対する回答の精度の向上が図れる。

【0073】請求項9に記載の発明によれば、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを格納したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第1のステップと、入力された質問文から検索結果集合と質問結果を判定する第2のステップと、前記検索結果集合および質問結果に従って、前記与えられた文書集合から関連文書集合を検索する第3のステップと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第4のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する回答結果として出力する第5のステ

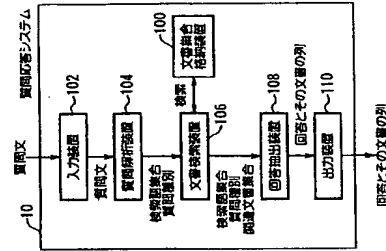
ップとをコンピュータに実行させる質問応答プログラムを記録媒体に記録したので、この記録媒体に記録した質問応答プログラムをコンピュータシステムに読み込ませ、実行することにより、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【図面の簡単な説明】

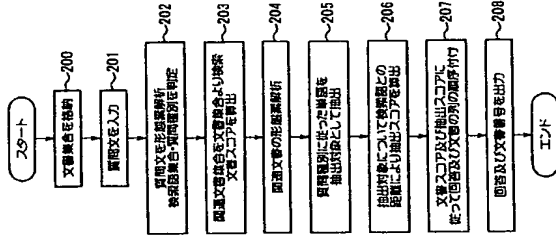
【図1】 本発明の実施の形態に係る質問応答システムの構成を示すブロック図。

【図2】 図1に示す質問応答システムの処理内容を示すフローチャート。

【図1】



【図2】



フロントページの続き

(72)発明者 平 博嗣
東京都千代田区大手町二丁目3番1号
本電信電話株式会社内

(72)発明者 平 裕
東京都千代田区大手町二丁目3番1号
本電信電話株式会社内

(72)発明者 加藤 信昭
東京都千代田区大手町二丁目3番1号
本電信電話株式会社内

Fターム(参考) 58075 N003 NK32 P724 P002 P074
58091 A111 A806 C402